EXAM 01 PRACTICE QUESTIONS

Solutions

Prof Amanda Luby

> **Note**
>
> These solutions are provided solely for Carleton College students enrolled in Amanda Luby's Stat120 course in Fall 2025. Dissemination of this solution to people who are not registered for this course is not permitted and will be considered grounds for an Academic Dishonesty report for all individuals involved in the giving and receiving of the solution.

This dataset gives education-related data for the 50 states and the District of Columbia. The variables are:

- `region`: West, Northeast, Midwest, South
- `pop`: Population, in 1,000's
- `verbal` and `math`: average SAT verbal and math scores
- `taken`: percent of students taking the SAT
- `noHS`: percent of population with no high school diploma
- `teachersPay`: median teacher salary, in 1,000's

a) Label each variable as categorical or quantitative

> **Note**
>
> - `region`: categorical
> - `pop`: quantitative
> - `verbal` and `math`: both quantitative
> - `taken`: quantitative
> - `noHS`: quantitative
> - `teachersPay`: quantitative

b) How many columns and rows are there in this dataset?

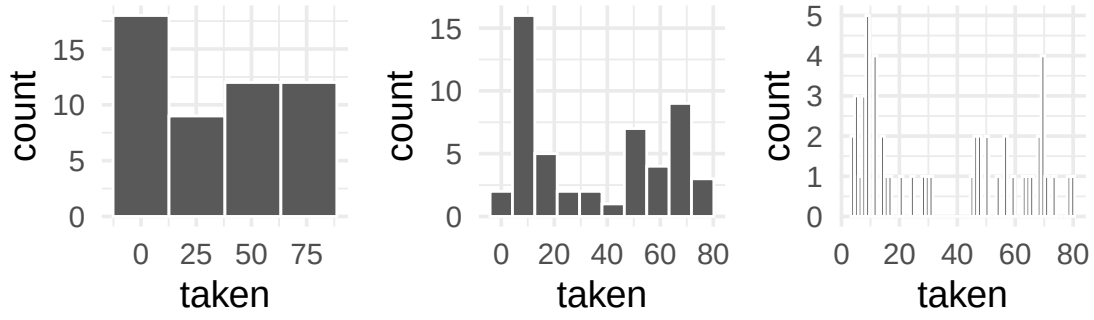
> **Note**
>
> 51 cases/rows by 8 variables/columns (including an ID variable)

c) Is this an experiment or observational study?

> **Note**
>
> Observational study

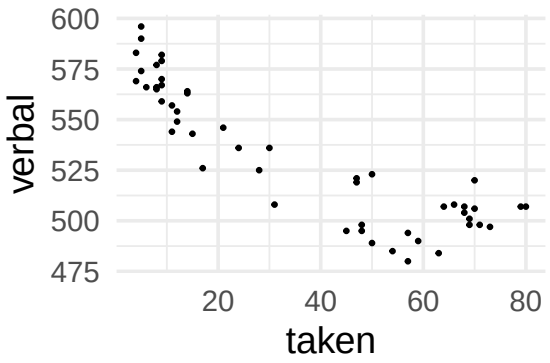
d) Below is the distribution of `taken` represented by 3 histograms. Choose the histogram that best represents the data and then describe the distribution.

> **Note**
>
> Middle histogram, the distribution is asymmetric/multimodal, skewed to the right, has a mode between 5-15, the mean will be higher than the mode, no outliers



e) Below is the scatterplot of `verbal` against `taken`. What do you think is closest to the actual correlation? Choose one: -1.1, -1, -.9, -.6, -.1, .1, .6, .9, 1, 1.1

> **Note**
>
> `-0.9 is closest to the actual correlation because the regression line would have a negative trend. The .88)`



f) What do the points in the *upper left* part of the graph represent? What about the *lower right*?

> **Note**
>
> Upper left- states where a small percentage of students took the SAT and got high verbal scores
> Lower right- states where a high percentage of students took the SAT and got lower verbal scores

g) Below is the linear regression model output for this data. Write out the linear regression equation and interpret the slope and intercept in context.

Note

$\hat{y} = 572.21 - 1.14x$
For a state where 0% of the students took the SAT, the average score is predicted to be 572 (this doesn't make sense though, there can be no scores for no students)
On average, states with a 1% increase in percent of students who took the SAT also tend to have a lower average score by 1.14 points.

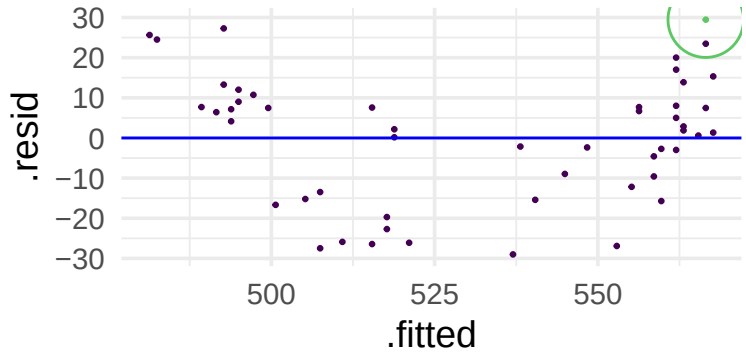
```
Call:
lm(formula = verbal ~ taken, data = sat)

Coefficients:
(Intercept)        taken
     572.21        -1.14
```

h) Below is the residual plot for this model. Do you have any concerns about the validity of the model?

Note

Yes, since the residual plot is showing some curvature, the linear model is not the best model for this set of data.



i) One point on the residual plot is colored in green and circled. Give the $\hat{y}$, ϵ, y, and x value for this data point.

Note

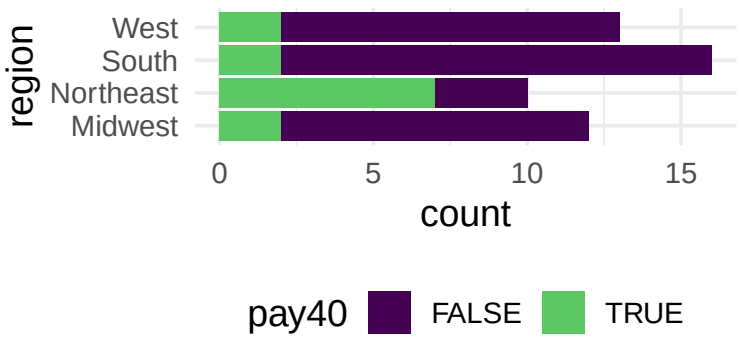
Epsilon = 30, y hat = 563, y = 593, x = 8.08

j) The mean of `teachersPay` is 35.89, the median is 35, and the standard deviation is 6.226. The distribution is approximately symmetric and bell-shaped. Approximately how many states have `teachersPay` between 29.66 and 42.12?

Note

This interval was found by using the 68/95/99 rule, which gives what percent of the data will be within 1/2/3 standard deviations of the mean respectively. Since it is a 95% interval, use 2 standard deviations.
lower interval bound = mean - 2 standard deviations = 35.89 - (2*6.226) = 23.44
Upper interval bound = mean + 2 standard deviations = 35.89 - (2*6.226) = 48.34
The result means that 95% of the data points have a value between 23.44 and 48.34.

j) The graph below displays `region` filled by a new variable, `pay40`: whether `teachersPay` is greater than 40.



k) Find the conditional distribution of `pay40` for the Northeast region

Note

Pay40	True	False
Northeast	7/10	3/10

l) Estimate the proportion of Midwest states with a median teacher pay less than 40.

Note

10/12

l) Find the marginal distribution of `pay40`

Note

True	False
13/51	38/51

m) Does there appear to be a relationship between `pay40` and `region`?

> **Note**
>
> Yes, there is a difference between the conditional distribution for the Northeast and the marginal distribution for all the data suggesting a relationship.

There will also be a few *conceptual* true/false or multiple choice questions that are *not* connected to the data analysis.